Title
Whole-proteome tree of life suggests a deep burst of organism diversity.

Journal
Proceedings of the National Academy of Sciences of the United States of America, 117(7)

ISSN
0027-8424

Authors
Choi, JaeJin
Kim, Sung-Hou

Publication Date
2020-02-01

DOI
10.1073/pnas.1915766117

Peer reviewed

Whole-proteome tree of life suggests a deep burst of organism diversity

JaeJin Choi[a,b,c] and Sung-Hou Kim[a,b,c,1]

[a]Department of Chemistry, University of California, Berkeley, CA 94720; [b]Center for Computational Biology, University of California, Berkeley, CA 94720; and [c]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Contributed by Sung-Hou Kim, December 11, 2019 (sent for review September 12, 2019; reviewed by Se-Ran Jun and Charles G. Kurland)

An organism tree of life (organism ToL) is a conceptual and metaphorical tree to capture a simplified narrative of the evolutionary course and kinship among the extant organisms. Such a tree cannot be experimentally validated but may be reconstructed based on characteristics associated with the organisms. Since the whole-genome sequence of an organism is, at present, the most comprehensive descriptor of the organism, a whole-genome sequence-based ToL can be an empirically derivable surrogate for the organism ToL. However, experimentally determining the whole-genome sequences of many diverse organisms was practically impossible until recently. We have constructed three types of ToLs for diversely sampled organisms using the sequences of whole genome, of whole transcriptome, and of whole proteome. Of the three, whole-proteome sequence-based ToL (whole-proteome ToL), constructed by applying information theory-based feature frequency profile method, an "alignment-free" method, gave the most topologically stable ToL. Here, we describe the main features of a whole-proteome ToL for 4,023 species with known complete or almost complete genome sequences on grouping and kinship among the groups at deep evolutionary levels. The ToL reveals 1) all extant organisms of this study can be grouped into 2 "Supergroups," 6 "Major Groups," or 35+ "Groups"; 2) the order of emergence of the "founders" of all of the groups may be assigned on an evolutionary progression scale; 3) all of the founders of the groups have emerged in a "deep burst" at the very beginning period near the root of the ToL—an explosive birth of life's diversity.

genome phylogeny | alignment-free | feature frequency profile | Jensen–Shannon divergence | genome tree

The tree of life (ToL) is a metaphorical tree that attempts to capture a simplified narrative of the evolutionary course and kinship among all living organisms, which cannot be validated experimentally.

Organism Tree of Life vs. Gene Tree of Life

The term "gene Tree of Life" (gene ToL) has been commonly used for the gene information-based ToLs constructed based on the sequence information, be it in DNA, RNA, or amino acid alphabets, of a set of selected genes or proteins coded by the genes. For decades, due to the technical difficulties of whole-genome sequencing, various gene ToLs have been used commonly as surrogates for the "organism ToL" despite the fact that gene ToLs most likely infer the evolutionary relationship of only the selected genes, not of the organisms.

Furthermore, there are various other intrinsic limitations and confounding issues associated with the construction and interpretation of gene ToLs (1). Many gene ToLs have been constructed based on different sets of the selected genes, new or increased number of extant organisms, and other inputs combined with various different gene-based analysis methods (1–10). They showed mostly good agreements on the clading of organism groups, but with varying degrees of disagreements on the branching orders and branching time of the groups, especially at deep tree branching levels. Thus, it became increasingly uncertain (1, 11, 12) whether gene ToLs are appropriate surrogates for the organism ToL. In addition, an important issue of rooting gene ToLs has not been well resolved and still is being debated (ref. 13 and references within).

These and other issues of gene ToLs highlight the need for alternative surrogates for the organism ToL built based on as completely different assumptions as possible from those of gene ToLs. A "genome ToL" (see below) constructed based on information theory (14) may provide an independent and alternative view of the organism ToL.

Genome ToL

Following the commonly used definition of gene ToL (see above), the term genome ToL is used in this study for the ToLs constructed based on the genomic information, be it DNA sequence of the whole genome, RNA sequence of whole transcriptome, or amino acid sequence of whole proteome, the latter two being derived from whole-genome sequence. (The term "whole" is to emphasize the entirety of the type of information derived from whole-genome sequences, rather than subjectively selected very small portions from whole-genome, whole-transcriptome, or whole-proteome information as in gene ToLs.)

The basic assumption is that the whole-sequence information, not the selected portion of the information, of an extant organism can be considered, at present, as the most comprehensive digital information of the organism for its survival and reproduction in its current environment and ecology. How to format such information

Significance

Tree of life (ToL) is a metaphorical tree that captures a simplified narrative of the evolutionary course and kinship among all living organisms of today. We have reconstructed a whole-proteome ToL for over 4,000 different extant species for which complete or near-complete genome sequences are available in public databases. The ToL suggests that 1) all extant organisms of this study can be grouped into 2 "Supergroups," 6 "Major Groups," or 35+ "Groups"; 2) the order of emergence of the "founders" of all the groups may be assigned on an evolutionary progression scale; and 3) all of the founders of the groups have emerged in a "deep burst" near the root of the ToL—an explosive birth of life's diversity.

Author contributions: S.-H.K. designed research; J.C. and S.-H.K. performed research; J.C. and S.-H.K. contributed new reagents/analytic tools; J.C. performed computer programming and execution; J.C. contributed extensive discussions; J.C. contributed computer-generated figures; J.C. and S.-H.K. designed figures; S.-H.K. contributed interpretations and implications of the results; J.C. and S.-H.K. analyzed data; and S.-H.K. wrote the paper.

Reviewers: S.-R.J., University of Arkansas for Medical Sciences; and C.G.K., Lund University.

The authors declare no competing interest.

Data deposition: The FFP programs for this study (2v.2.1) written in GCC(g++) are available in GitHub (https://github.com/jaejinchoi/FFP).

[1]To whom correspondence may be addressed. Email: sunghou@berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915766117/-/DCSupplemental.

First published February 4, 2020.

system ("descriptor") and quantitatively compare a pair of the systems ("distance measure" for the degree of difference, dissimilarity, or divergence) are well developed in information theory, not only for digitally encoded electronic signals and images, but also for natural language systems, such as books and documents (15).

Descriptor. In this study, we use the descriptor of feature frequency profile (FFP) method (ref. 16; see *FFP Method* in *Materials and Methods*) to describe the whole-genome sequence information of an organism in DNA, RNA, or amino acid alphabets. Briefly, a feature in FFP is an adaptation of "*n*-grams" or "*k*-mers" used to describe a sentence, a paragraph, a chapter, or a whole book (17, 18) in information theory and computational linguistics, where an *n*-gram is a string of *n* alphabets plus space and delimiter such as comma, period, etc. For this study, we treat the genome information (whole genome, whole transcriptome, or whole proteome) of an organism as a book of alphabets without spaces and delimiters, and represent the "genome book" by its FFP, which is the collection of all unique *n*-grams and their frequencies in the book. Thus, for a whole genome, the features and their frequencies in the genome are analogous to "characters" and "character states," respectively, for the genome. It is important to note that such FFP of a sequence has all of the information necessary to reconstruct the original sequence. The "optimal *n*" of the *n*-grams, the most critical parameter, for the construction of a sequence-based genome ToL can be empirically obtained under a given criterion (see *Choice of the "Best" Descriptor and the "Optimal" Feature Length for a Whole-Proteome ToL* in *Materials and Methods*). For the criterion of the most topologically stable ToL, it usually ranges, depending on the size of the sample and the types of information ("alphabets" of genome information), between 10 and 15 or longer for whole-proteome sequence and between 20 and 26 for whole-genome or whole-transcriptome sequences. For this study, we found that the whole-proteome sequence with the optimal feature length of 12 amino acids or longer yields the most topologically stable genome ToL (*SI Appendix*, Fig. S1). For short, this whole-proteome-based genome ToL will be referred to as "whole-proteome ToL."

"Distance Measure." As for the measure to estimate the degree of difference between two FFPs, we use Jensen–Shannon divergence (JSD) (19) (for details, see *JSD and Cumulative Branch Length* in *Materials and Methods*), an information-theoretical function for estimating the extent of differences or divergence between two linear information systems. Such divergences for all pairs of organisms, then, can be used to assemble the "divergence distance matrix" needed to build a genome ToL.

Outgroup. The descriptor and the distance measure used in the FFP method provide another important advantage for constructing the genome sequence of an "artificial (faux) organism" that may be used as an outgroup member of a ToL. An artificial genome, transcriptome, or proteome can be constructed to have the same size and composition of genomic alphabets as one of the real extant organism in the study population, but the sequence of the alphabets are shuffled within the real organism, such that it has presumably no information for sustaining and reproducing life. Since the FFP method is one of the "alignment-free" methods, which do not require multiple sequence alignment of long stretches of sequences common among all members of the study population, such artificial organisms have been used successfully as members of an outgroup in constructing the rooted whole-genome trees for prokaryotes (20) and fungi (21).

Pool of "Founding Ancestors" vs. Common Ancestor. Recent observations prompted us to revisit the meaning of internal nodes of a ToL for this study: Whole-genome sequences of a very large number of *Homo sapiens* and *Escherichia coli* species have been experimentally determined in the last decade. They revealed that the extent of the genomic divergence and variation among the members in a species are very broad even after a short period of evolutionary time (22–24) and even under a constant environment in the case of *E. coli* (24). Thus, an internal node can be considered, in this study, as a pool of founding ancestors (FAs) with wide genomic diversity, from which divergent founders (small subpopulations of the pool) for the new groups ("founder effect") emerge or "sampled/selected," under, for example, drastic changes in various local environment and ecology. This "mosaic" feature of the internal node is conceptually different from the "clonal" feature of the node as a common ancestor, from which two or more descendants with high genomic similarity branch out (*SI Appendix*, Fig. S2).

Objective. Extending our earlier experiences with constructing the rooted whole-proteome trees of prokaryotes (20) and fungi (21), we constructed a "rooted whole-proteome ToL" of 4,023 species, using all predicted protein sequences encoded by all predicted genes of each organisms, ranging from the smallest proteome of 253 proteins of *Candidatus Portiera aleyrodidarum*, a bacterial symbiont of a whiteflies, to 112,718 proteins of *Brassica napus*, a land plant. The ToL reveals some unexpected features and notable differences compared to the existing gene ToLs. It is hoped that these differences stimulate additional and/or alternative narratives for some of the important aspects of the organism ToL.

EVOLUTION

Results

In this section, we present our observations of the features in our whole-proteome ToL. Associated implications and narratives will be presented in *Discussion*. To highlight the similarities and differences of the features of the whole-proteome ToL from those of gene ToLs, we present our results from two viewpoints: 1) identification of large groups and the topological relationship among the groups showing the order of emergence of each large group, and 2) relative magnitude in cumulative branch lengths among the founders of all of the groups to estimate the relative extent of evolutionary progression at which the founders emerged and eventually evolved toward the respective extant organism groups. Since the grouping and the branching order of the groups in the gene ToLs do not always agree with those in our whole-proteome ToL, we use the following descriptions for the groups at various branching levels in our ToL: "Supergroup," "Major Group," "Group," and "Subgroup." The generic labels of the groups in the ToL are assigned and the corresponding taxon names from National Center for Biotechnology Information (NCBI) (see *Sources and Selection of Proteome Sequences and Taxonomic Names* in *Materials and Methods*), which are mostly based on the clading pattern in the gene trees and characteristic phenotypes at the time of naming, are also listed for comparison.

Grouping of Extant Organisms: Two Supergroups, Six Major Groups, or 35+ Groups. Fig. 1 shows that, at the deepest level, two Supergroups emerge as indicated by the two-colored inner circular band: the red-colored portion corresponds to Prokarya ["Akarya" (25, 26) may be a more appropriate name, because the founders of Prokarya do not emerge before the founders of Eukarya in our ToL (also see Fig. 2)] and the blue colored portion to Eukarya. At the next level, six Major Groups emerge as indicated in the outer circular band by six colors. Finally, 35 Groups emerge as indicated by the small circles with their whole-proteome ToL labels (next to the circles) and corresponding scientific or common names used in NCBI database (outside of the circular bands).

The membership of each of all eukaryotic Major Groups and Groups in our ToL coincide with those of the groups identified by NCBI taxonomic names at a phylum (P), a class (C), or an order (O) level with a few exceptions marked by asterisks before

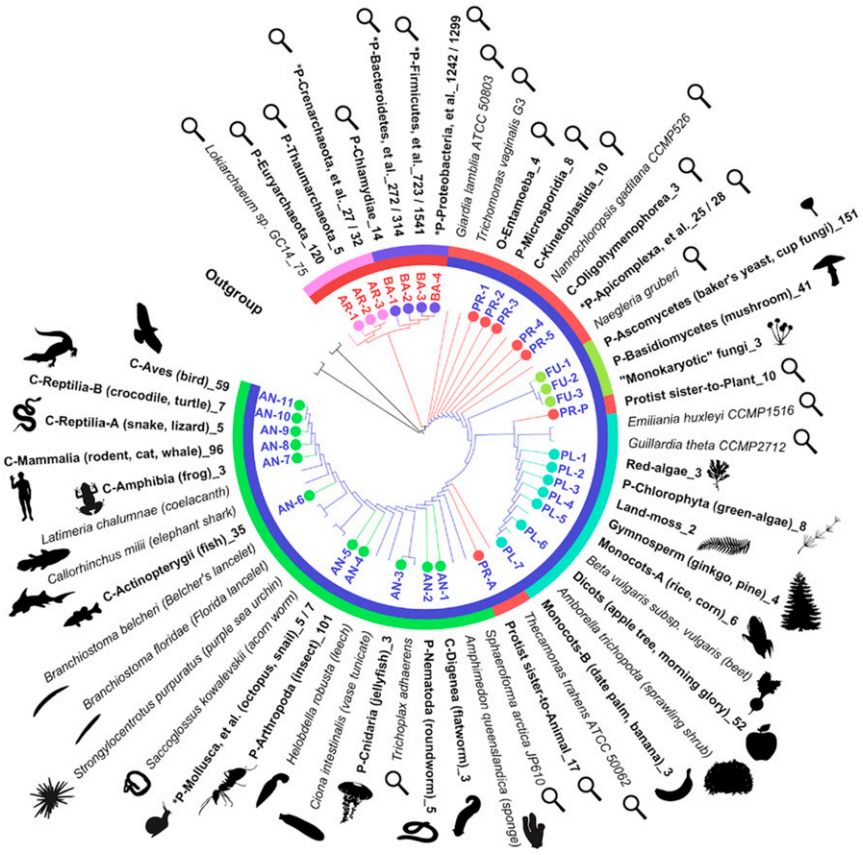

Fig. 1. Topology and branching order of a whole-proteome ToL. A Circos ITOL (47) diagram highlighting the topology and branching order of the whole-proteome ToL of all "collapsed" clades (small colored circles) as well as most of singletons (in italics) in the study population. The branch lengths are ignored. The inner ring has two colors for two Supergroups: Red corresponds to Prokarya (Akarya) and blue to Eukarya. The outer ring has six colors for six Major Groups: pink for AR (Archaea), blue for BA (Bacteria), orange for PR (Protists in three types: PR-1~5, PR-P, and PR-A), pale green for FU (Fungi), cyan for PL (Plants), and green for AN (Animals). The labels for 35 Groups (3 for AR, 4 for BA, 7 for PR in three Groups, 3 for FU, 7 for PL, and 11 for AN) are shown next to the small circles, and their corresponding scientific names according to NCBI taxonomy (preceded by taxonomic level of phylum [P], class [C], or order [O]) are shown outside of the outer ring followed by the number of samples in this study. Common names and silhouettes of one or more examples are also shown. The symbol of a magnifying glass represents microbial organisms not visible by human eyes. Asterisks are for the clades with "mixed" names: For example, the group "*P-Bacteroidetes _272/314" consists of 314 members of several bacterial phyla, of which 272 are Bacteroidetes, the phylum name of the majority in the group. The visualization of ToL was made using ITOL (47).

the taxonomic names (see Fig. 1 legend). A notable exception is three Groups of Major Group Protists (see *Three Types of Major Group Protists* in *Phylogeny of Supergroup Eukarya* below).

For prokaryotes (akaryotes), this is also the case at Major Group level, but not at Group level [see *Phylogeny of Supergroup Prokarya (Akarya)* below]. However, as described below, there are significant differences in the topological relationship among the groups, relative branch lengths, and branching order associated with the groups between our whole-proteome ToL and gene ToLs (see below).

First Emergence of the "Founders" of Supergroup Prokarya (Akarya) and Supergroup Eukarya. Fig. 2 shows that both the founder (see *Pool of "Founding Ancestors" vs. Common Ancestor* and *SI Appendix*, Fig. S2) of the Supergroup Prokarya (Akarya) and the founder of the Supergroup Eukarya emerge first (as in "Eukarya early" model) from the pool of last "terrestrial (Earth-bound)" founding ancestors (LTFAs) [analogous to LUCA (last universal common ancestor), but emphasizing the aspects of terrestrial (earthly) ancestors and a pool of FAs of the first internal node of our ToL]. Then, the founders of Major Groups Archaea and Bacteria and those of all eukaryotic Major Groups (Protists, Fungi, Plants, and Animals) emerge from their respective Supergroup FA pools. This is in contrast to the branching out of Archaea and Bacteria from LUCA first, followed by branching out of Eukarya from Archaea accompanied by one or more steps of fusion events between some members of Bacteria and Archaea (as in "Eukarya late" model), according to most of the recent gene ToLs (4, 5, 7–11, 27, 28, 29). However, it is in agreement with our earlier rooted whole-proteome tree of prokaryotes constructed using the whole-proteome sequences of 884 prokaryotes plus 16 unicellular eukaryotes (20), and a ToL reconstructed using the contents of coding sequences for protein "fold-domain superfamily" and rooted in a way very different (ref. 25 and 26 and references within) from most gene ToLs and our whole-proteome ToL.

"Deep Burst" of the Founders of all Major Groups. Fig. 2 shows "cumulative genome information divergence" (CGD) values, which are, in this study, the scaled cumulative branch lengths along the lineage for all of the FA nodes (internal nodes) of the five Major Groups and three protist Groups. Thus, CGD values represent the extent of evolutionary progression and are scaled from 0.00 for the root of the ToL (at the start of evolution) to an average of 100.00 for the extant organisms (at the present state of evolution) (see *Chronological Timescale vs. Progression Scale* in *Discussion*). The figure also highlights that the founders for all Major Groups in the study population have emerged abruptly

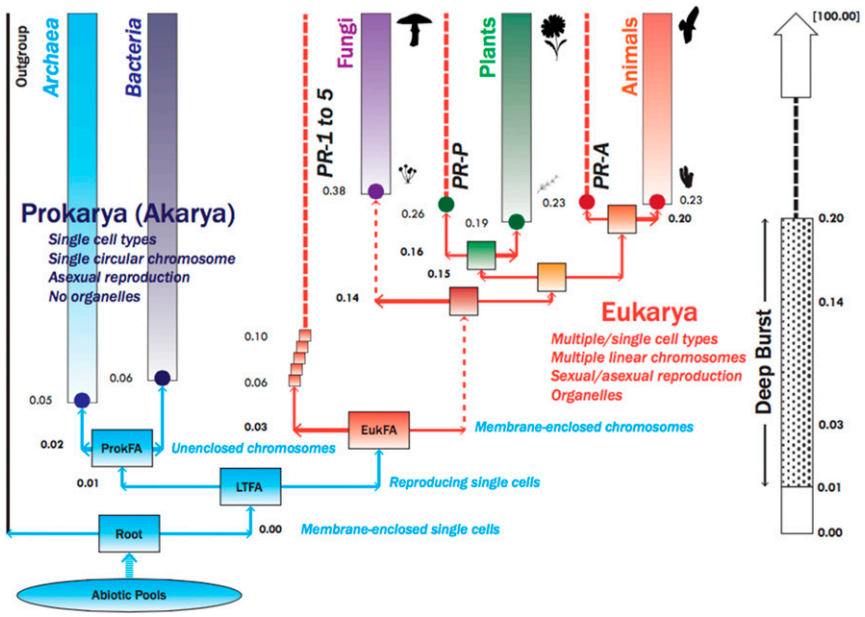

Fig. 2. Simplified whole-proteome ToL at the deepest level. All extant organisms in this study are grouped into five Major Groups, as shown as five columns (corresponding to Archaea, Bacteria, Fungi, Plants, and Animals), and a paraphyletic protist "Group" represented by three thick dotted red columns (corresponding to three groups of protists, labeled as PR-1~5, PR-P, and PR-A, where P is for "sister to Plants" and A for "sister to Animals"). For simplicity, singletons, the organisms that do not belong to any named groups, are not shown. It also shows the scaled cumulative branch lengths to internal nodes (rectangles). Each small circle represents the internal node of the clade containing all extant members of the clade of a Major Group, PR-P or PR-A subgroup, and each rectangular box presents a pool of the founding ancestors (FAs) from which one or more founders emerge (or are selected) to evolve to become a node containing an extant organism or the "seed" for the next FA pool (*SI Appendix*, Fig. S2). The bold number next to each horizontal arrow is the cumulative genome information divergence (CGD) value (scaled cumulative branch length from the root), at which a small subset of the founder(s) of the respective Major Group emerged, and the plain number next to each circle corresponds to the CGD value of the internal node of a clade containing all the extant members of the Major Group. The silhouettes of one of the early and one of late emerged organisms of each Major Group among the study population are shown: a small member of Ascomycota and a mushroom for Fungi, an algae and a flowering land plant for Plants, and a sponge and a bird for Animals, respectively. The broad arrow on *Right* is to schematically show that all "founders" of the six Major Groups emerged in a deep burst within a very short range of CGD between 0.01 and 0.20 on the progression scale of evolution (see *Chronological Timescale vs. Progression Scale* in *Discussion*), which is scaled to an average of 100.00 for the extant organisms. LTFA, the pool of last terrestrial (Earth-bound) founding ancestors of replicating cells; root, the root of the whole-proteome ToL, the last pool of nonreplicating cells of diverse contents; ProkFA, the pool of Prokarya founding ancestors; EukFA, the pool of Eukarya founding ancestors. The very first step between "abiotic pools" and the root may include one or more (indicated by a multistepped arrow) catastrophic events of death of all previous life forms.

during a three-staged deep burst of genomic divergence near the very beginning of the root of the whole-proteome ToL: 1) At the first stage (CGD of 0.01), the FAs of the two Supergroups, Prokarya (Akarya) and Eukarya, emerge from LTFA; 2) in the second stage (CGD of 0.02 to 0.03), the FAs of three types of unicellular organism groups emerge: Archaea and Bacteria emerge from the pool of Prokarya FAs, and the founders of the unicellular protist Groups (PR-1 to PR-5) emerge from the pool of Eukarya FAs; and 3) in the third stage (CGD of 0.14 to 0.20), the founders of three Major Groups corresponding to Fungus, Plant, and Animal groups plus two Protist groups (PR-P and PR-A) emerge. Thus, the abrupt emergence of the founders of all five Major Groups plus a protist Major Group (composed of three types of protists) occurred during the deep burst within a very short range between 0.01 and 0.20 on the progression (i.e., CGD) scale of evolution, followed by "relatively gradual (non-abrupt)" evolution of the Major Groups a long period corresponding to CGD value of 99.80 (see *Chronological Timescale vs. Progression Scale* in *Discussion*).

Emergence and Divergence of the Founders of 35+ Groups. Figs. 3 and 4 show that each founder of 35+ Groups emerged during a period corresponding to CGD value between 0.05 (emergence of the founder of Archaea in Fig. 2) and 27.62 (emergence of the founders of extant birds and crocodile/turtle Groups in Fig. 3), suggesting that, depending on the Group, 99.95 to 72.38 in CGD scale account for the "relatively gradual" (not "bursting") evolution toward the extant organisms.

Phylogeny of Supergroup Eukarya.

Branching orders within three Major Groups. Fig. 3 shows the order of the emergence of the founders of all eukaryotic Major Groups. The branching order of the three Major Groups (Fungi, Plants, and Animals) differs from those of the gene ToLs: Almost all recent gene ToLs show Fungi as the sister clade of Animal clade (for a recent review, see ref. 27), but in our whole-proteome ToL, Major Group Fungi is sister to the combined group of the Plant and Animal Major Groups plus their respective Protist sister Groups, PR-P and PR-A.

For Major Group Fungi, as reported earlier (21), the founders of all three Groups of Fungi corresponding to Ascomycota, Basidiomycota, and "Monokarya" ("non-Dikarya") emerged within a small CGD range of 0.38 to 0.44, of which the founders of the Ascomycota appears first at CGD of 0.38, around a similar value of CGD when the founders of red and green algae of Plants and of invertebrates of Animals emerged.

In Major Group Plants, the order of emergence of the founders starts with those of marine plants, such as red algae and green algae. After a large jump of CGD value, the founders of nonflowering land plants such as spore-forming ferns and land mosses emerged, then "naked" seed-forming gymnosperms such as gingko and pines, followed by "enclosed" seed-forming

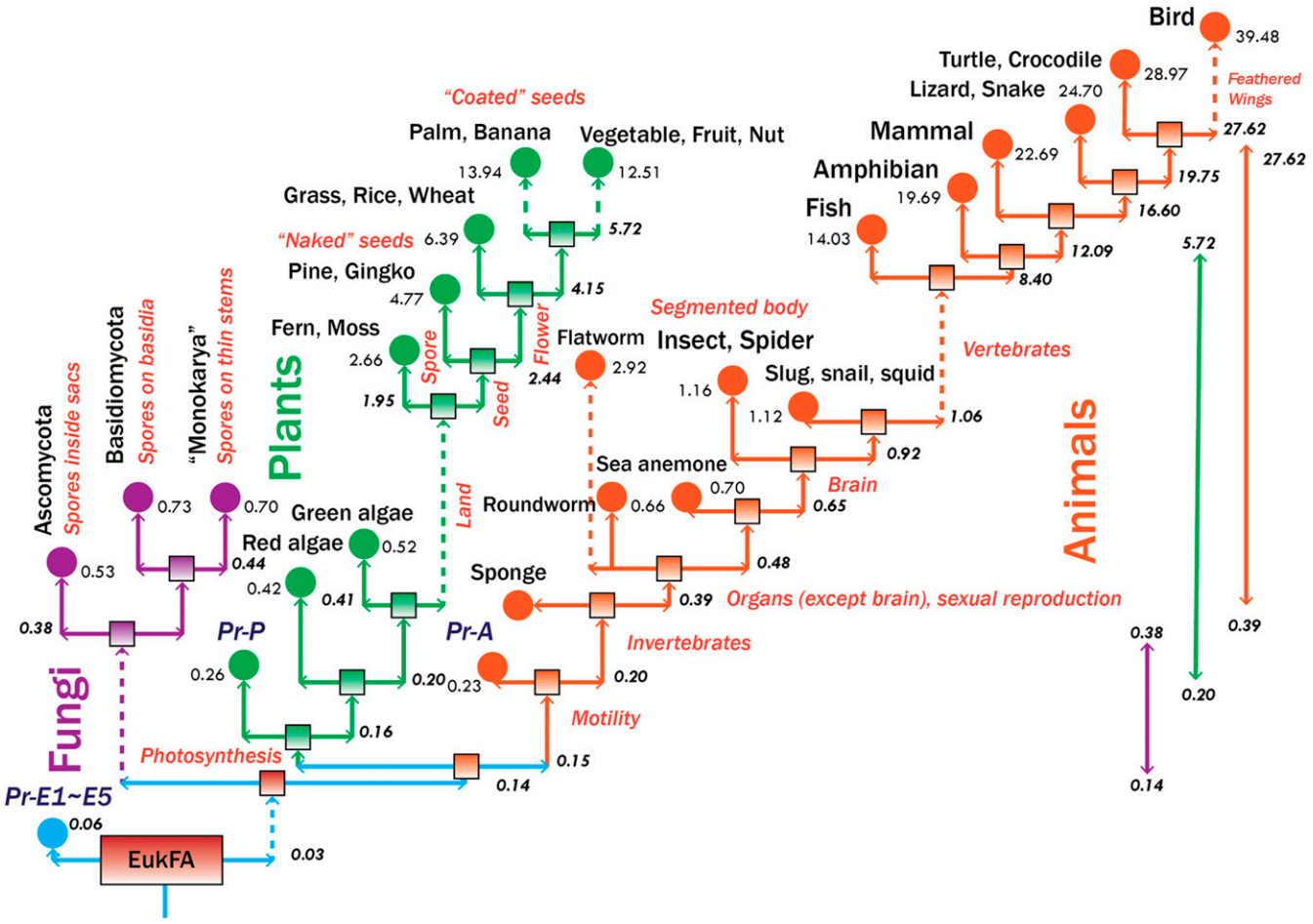

Fig. 3. Simplified eukaryotic portion of the whole-proteome ToL at Group level. The portion of the whole-proteome ToL corresponding to all eukaryotes is shown at the tree-branching level of Groups. For simplicity, singletons (that do not belong to any named groups) are not shown. Also not shown are all leaf nodes and their branches to their common ancestor nodes. Rectangles and circles as well as the bold number next to each horizontal arrow and the plain number next to each circle have the same meanings as those in Fig. 2. The dotted vertical arrows are to indicate that they are arbitrarily shortened to accommodate large jumps of CGD values within a limited space of the figure. The range of the CGD values, within which all of the founders of the extant organisms of each Major Group have emerged, are shown on the *Right*. The CGD values of each colored Major Group are not scaled to those of the others.

flowering land plants of angiosperms such as monocot and dicot plants.

In Major Group Animals, the order of the emergence of the founders starts with those of invertebrates such as sponges, worms, Cnidaria, arthropods, and Mollusca. Then, after a big jump of CGD, the founders of vertebrates such as fishes, amphibians, mammals, and reptile-A (snakes, lizards, etc.), emerge, and finally, the sister pair of the founders of reptile-B (crocodiles, alligators, turtles, etc.) and birds, of which the extant birds with feathered wings emerge after another big jump of CGD.

All of the founders of eukaryotic Groups emerged in a very wide range of CGD between 0.20 and 27.48 (Fig. 3). This is in drastic contrast to the very small CGD range of 0.06 and 0.12 in which all founders of prokaryotic Groups emerged [see *Phylogeny of Supergroup Prokarya (Akarya)* below].

Three types of Major Group protists. Protists are currently defined as unicellular eukaryotic organisms that are not fungi, plants, or animals. Despite the paucity of the whole-genome information of protists (30), especially of a large population of nonparasitic protists, the whole-proteome ToL suggests that there are at least seven smaller groups of protists that can be divided into three types as mentioned earlier. Fig. 2 shows that the protist Subgroup PR-P is the sister clade to Major Group Plants (red and green algae plus land plants), and most of them belong to one of two categories: photosynthetic protists, such as microalgae, diatoms, phytoplankton, or water molds parasitic to plants. Another protist Subgroup, PR-A, is a larger group including many protists with a wide range of phenotypes and is the sister group to Major Group Animals. Most of them are motile and can be grouped with slime molds, amoeboids, choanoflagellates, and others, or with parasites to cattle and poultry (Fig. 3). The rest of protists form five small subgroups (labeled as PR-1 to PR-5 in Figs. 1–3), which do not form a larger clade together but emerge successively from the pool of Eukarya FA. Most of them are parasitic to animals with varying host specificity. The founders of these protist Subgroups emerged much earlier than those of PR-P and PR-A as implied by smaller CGDs, and they correspond to Entamoeba, Microsporidia, Kinetoplastida, Oligohymenophorea, and Apicomplexa, respectively.

Phylogeny of Supergroup Prokarya (Akarya). Deep branching pattern of Supergroup Prokarya (Akarya) is much more "collapsed" beyond the Major Group level compared to that in Supergroup Eukarya. Fig. 4 shows a simplified prokaryotic portion of the whole-proteome ToL. At the deepest level of Supergroup Prokarya (CGD of 0.02 from the root), the founders of two Major Groups, corresponding to Archaea and Bacteria, emerge, followed by the emergence of the founders of three Groups of Archaea (AR-1, -2, and -3 corresponding approximately to the

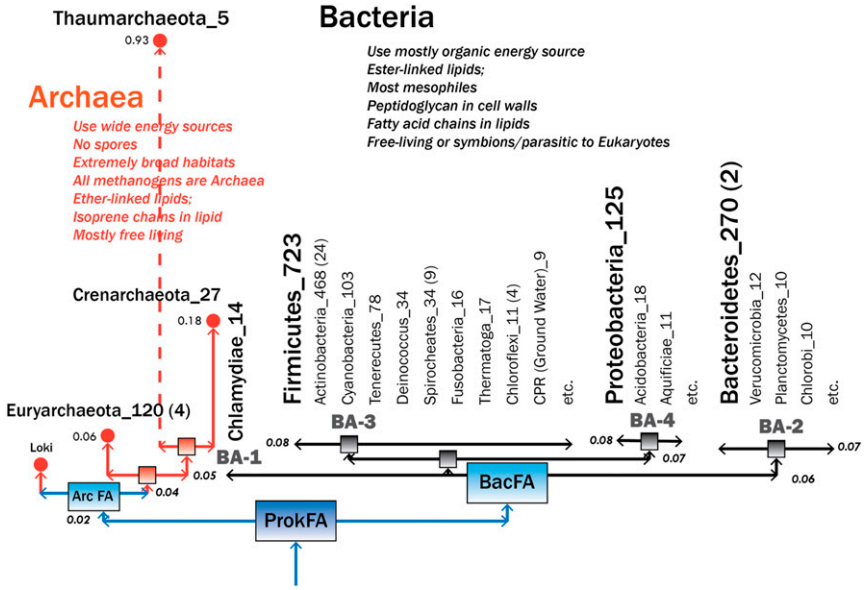

Fig. 4. Simplified prokaryotic portion of the whole-proteome ToL at Group level. The portion of the whole-proteome ToL for prokaryotes are shown at Group (mostly phylum) level, where, for simplicity, Subgroups with small sample sizes and singletons are not shown. Major Group Bacteria is divided into four Groups, BA-1 to BA-4, where each Group consists of one or many Subgroups. The boldfaced name in each Group has the largest sample size among the members of the Group. Number after each NCBI taxonomic name refers to the sample size of the majority clade, followed by a number in parentheses referring to the size of minority clade away from the majority clade. Interestingly, the founders of all of the named Groups of Bacteria emerged within a very small range of CGD values of 0.06 to 0.08, in a drastic contrast to a much larger range of 0.20 to 27.62 for those of all eukaryotes. (The numbers next to horizontal lines and circles have the same meanings as those in Fig. 2.) Thus, the order of the emergence of all of the named bacterial Groups and Subgroups may be less reliable compared to that of eukaryotes. For simplicity, the branching order among the Subgroups is not shown.

extant Euryarchaeota, Thaumarchaeota, and Crenarchaeota) and four large bacterial Groups (BA-1 to BA-4), with no obvious distinguishing characteristics, within a very small CGD range of 0.05 to 0.06 [Fig. 4; see also *On the Phylogeny of Supergroup Prokarya (Akarya)* in *Discussion*]. Of the four Groups, BA-1 has only one member, *Chlamydia*, obligate intracellular parasites, and is basal to the remaining three. BA-3 is the largest and is very similar to an unranked "Terrabacteria Supergroup," whose habitat is strictly "nonmarine" (e.g., soil or rock on land; freshwater of lakes, rivers, and springs), or if their host is a nonmarine species (31). Members of this Group include those resistant to environmental hazards such as UV radiation, desiccation, and high salinity, as well as those that do oxygenic photosynthesis. The remaining two Groups have no proposed Supergroup names yet.

Among the four bacterial Groups in this study population, there are more than 30 bacterial Subgroups corresponding to the groups with NCBI taxon names at the phylum level. The founders of all of them emerged within a small CGD range of 0.06 and 0.12 (Fig. 4), thus making it less certain about the resolution of the branching order not only among the four Groups but also among many Subgroups within each Group. This is a drastic contrast to a very large CGD range between 0.20 and 27.48 in which all founders of eukaryotic Groups emerged (see *Phylogeny of Supergroup Eukarya* above).

Phylogenic Positions of Recently Discovered Groups. Candidate phyla radiation (CPR) group is a very large group of small bacteria with relatively small genomes. Most of them are found in diverse environments, including groundwater, and are symbiotic with other microbes in their community; thus, it is difficult to culture their representative members for whole-genome sequencing. An extensive metagenomic study and gene tree construction using 16 ribosomal protein genes revealed that the CPR group members form a "Supergroup," which is "well-separated from all other Bacteria" (8, 9). Despite its vast population size, only eight genome sequences are available at present in public databases. In our whole-proteome ToL, these small samples form a sister clade to Tenericutes, a member of Major Group BA-3 of Bacteria (Fig. 4). More full-genome sequences of other CPR group members may resolve this apparent discrepancy of grouping and interpretation.

Hemimastigophora group is one of the unicellular eukaryotic protist groups with uncertain phylogenic assignment due to the absence of genomic sequence information. The members of the group are mostly free-living and predatory to other microbes in soil, sediment, water column, and soft water environment, and have highly distinctive morphology. Based on recent studies of transcriptome sequences of two members of the group, an unrooted gene tree using 351 common genes was constructed and reported that Hemimastigophora forms a new "supra-kingdom" at the basal position of the Animal kingdom (32). Our whole-proteome ToL is showing the group as a member of the protist Subgroup of PR-A, which is the basal or sister clade to Major Group Animals.

Phylogenic Positions of "Singletons." Fig. 1 and *SI Appendix*, Fig. S3 show our ToL grouped at phylum, class, or order level: the former as a circular topological ToL ignoring the branch lengths, and the latter as a linear ToL with all of the cumulative branch lengths shown for internal nodes from which all of the founders of the groups emerge. In both figures, most of the "singletons" are included. Singleton is defined, for this study population, as an organism that does not find a closest neighbor of the same group name. Many of them are at basal/sister positions to larger groups, suggesting possible speculative evolutionary roles, accompanied with "accumulative" or/and "reductive" evolution. In one case, possible reassignment of group affiliation is suggested:

1) *Lokiarchaeum* sp. as a basal organism to all members of Major Group Archaea, rather than a member of the sister group to Euryarchaeota (27).

2) *Giardia lamblia* and *Trichomonas vaginalis* at the basal or sister position to Supergroup Eukarya.
3) *Naegleria gruberi* at the basal position of three Major Groups of Fungi, Plants, and Animals plus two protist Subgroups, PR-P and PR-A.
4) *Emiliania huxleyi* and *Guillardia theta* as a nearest-neighbor pair at the basal position of Major Group Plants.
5) *Thecamonas trahens* at the basal position of Major Group Animals plus the protist Subgroup PR-A.
6) *Sphaeroforma arctica* and *Amphimedon queensiandica* at the basal positions of Major Group Animals (33).
7) *Trichoplax adhaerens* at the basal/sister position to Cnidaria clade (slug, snail, squid) of Major Group Animals.
8) *Ciona intestinalis* and *Hellopdella robusta* emerging between Cnidaria clade and Arthropoda clade (insect, spider) of Major Group Animals.
9) *Branchiostomas* emerging before Group Fish of Major Group Animals.
10) *Callorhinchus milii* and *Latimeria chalumnae* emerging between Group Fish and Group Amphibian of Major Group Animals.

The genome sequences of more organisms that are close relatives of these singletons are needed to confirm or refute these speculations.

Discussion

Chronological Timescale vs. Progression Scale. There is no known measure to estimate the chronological evolutionary time along the evolutionary lineage of an extant organism, especially in deep evolutionary period when no fossils are available. However, the degree of evolutionary progression from the root of our ToL to a given internal node (FA node) can be represented by the scaled cumulative branch length (i.e., CGD value) of the internal node (Fig. 2). Although the chronological evolutionary timescale is different from the evolutionary progression scale, both scales have the same directional arrow starting from the root to the extant organisms; thus, the ranking order (branching order) of the emergence of the founders along the lineage in both scales are the same.

Deep-Burst Model vs. Other Models for Evolution of Organisms. As mentioned earlier in *Results*, the founders of all Major Groups emerged by the time corresponding to 0.20 on CGD scale (Fig. 2). Such explosive birth model of Deep Burst has some similarities to various aspects of earlier models of evolution of life: 1) The nontree-like "Biological Big Bang" (BBB) hypothesis (34), especially the second of the three BBBs, is similar to the emergence of the first stage of the deep burst, and 2) the unresolved "bush-like tree" model (35), especially the "collapsed" aspect of deep branches at the Group level of Prokarya (Akarya). Both models are inferred from various analyses of gene ToLs and they appear to correspond to one of the three stages of the deep burst, all of which occurred during the period corresponding to 0.20 on CGD scale.

The remaining 99.80 of CGD scale (the evolutionary progression scale) accounts for presumed multiple "relatively" gradual (less abrupt) evolutionary steps, such as multiple cycles of emergence of new founders and gradual divergence of their genomes. Such gradual steps have similarity to "punctuated equilibrium" model (36) inferred from paleontological analyses of fossils.

On the Phylogeny of Group Animal. The sequence of the emergence of the founders of all named Groups, as shown in Fig. 3, agrees with that of most of the gene ToLs as well as fossil data except that of mammals and birds. Many gene ToLs and fossil data suggest mammals emerging after emergence of birds and reptiles. In our ToL, the founder of Group Mammal is sister to those of the joined Groups of birds and two types of reptiles. Such sisterhood was also detected in some gene ToLs (e.g., ref. 4 and 5). Furthermore, Fig. 3 also shows that the FA of the extant mammal Group of this study population, indicated as a circle with CGD of 22.69 in Fig. 3, emerged earlier than those of the extant bird Group and the extant reptile Groups. This is "counterintuitive," although one can imagine a narrative that all premammal birds and reptiles did not survive certain mass extinction event(s), but some of those are detected only as fossils.

On the Phylogeny of Supergroup Prokarya (Akarya). For all eukaryotic organisms in the study, the membership of the organisms in each of the 21 Groups (3 fungus Groups, 7 plant Groups, and 11 animal Groups; 3 protist Groups are not counted) agrees well with that of the organisms in most gene ToLs at phylum/class names (Fig. 3). This is also true for most prokaryotic groups (about 30 groups at phylum level) with some exceptions, where one or more small minority of a group does not clade with their respective majority group (Fig. 4). There may be many possible reasons for the minority "discrepancies," but they are expected to be resolved once more whole-genome sequences of diverse members of the small minorities become available.

Similar Composition of Each Clade but Different Branching Order of the Clades. We found it surprising that, although the branching order of various groups are significantly different between gene ToLs and our whole-proteome Tol, the membership of each group in most groups is the same not only at all Supergroup and Major Group levels, but also at most Group level (corresponding mostly to phylum level). This is not expected because the membership of each group is assigned by the clading pattern, which is determined by two completely different methods. One possible explanation is that this surprising observation is the consequence that, after the deep burst, when the founders of all groups emerged, all of the organisms within each group evolved relatively "isolated" from other groups for the evolutionary time corresponding to most of CGD scale (72.4 to 99.9). We attribute the differences in the branching order of the groups to the fact that the branch lengths are calculated by two totally different distance measures applied on two completely different descriptors for the organisms.

A Narrative. Fig. 2 is a simplified whole-proteome ToL, which suggests a possible narrative of the evolutionary course and kinship among the large groups of extant organisms in this study. Briefly, it suggests that 1) the replicating cells of LTFAs may have emerged from diverse nonreplicating cells formed by random packaging of various assortments of molecules, including stable smaller circular DNAs ("pre-Akaryan" cells) and prepacked larger linear DNAs ("pre-Eukaryan" cells), by self-assembling and fusing of membranes; 2) two Supergroups (Prokarya, also known as Akarya, and Eukarya) may have emerged from the LTFAs of diverse size and content, not from the last clonal common ancestor; and 3) the founders of all of the Major Groups (Archaea, Bacteria, Protists, Fungi, Plants, and Animals) emerged during a deep burst, near the root of the ToL—an explosive birth of life's diversity. All emergences may have occurred by "selection" under critical and drastic environmental changes.

This narrative is contrasted, in *SI Appendix*, Fig. S4, to those of most current gene ToLs. The figure schematically emphasizes the differences in grouping, branching order of the groups, and the nature of the internal nodes ("mosaic population" of ancestors vs. clonal ancestor) between the whole-proteome ToL and gene ToLs.

Materials and Methods

Sources and Selection of Proteome Sequences and Taxonomic Names. All publicly available proteome sequences used in this study are obtained from the NCBI. We downloaded the proteome sequences for 691 eukaryotes and 3,317 prokaryotes from NCBI RefSeq DB using NCBI FTP depository (ref. 37; as

of July 2017, our project start time). Proteome sequences derived from all organelles were excluded from this study. In addition, we included the proteome sequences of nine prokaryotes (as of August 2017): Lokiarchaeum (27) and eight CPR groundwater bacteria (8, 9) from NCBI GenBank. We also included six eukaryotes: four gymnosperms (*Ginkgo biloba*, *Pinus lambertiana*, *Pinus taeda*, and *Pseudotuga menziesi*) from Treegenesdb (ref. 38; as of June 2018); and two Hemimastigotes (*Hemimastix kukwesjijk* and *Spironema* cf Multiciliatum) (ref. 32; as of January 2019) derived from the transcriptome using "TransDecoder."

Thus, the total of 4,023 proteome sequences form the population of this study.

Proteome sequences not included in our study are those derived from whole-genome sequences assembled with "low" completeness based on two criteria: 1) the genome assembly level indicated by NCBI as "contig" or lower (i.e., we selected those with the assembly levels of "scaffold," "chromosome," or "complete genome"), and 2) the proteome size smaller than the smallest proteome size among highly assembled genomes of eukaryotes and prokaryotes, respectively. For the minimum proteome size threshold for eukaryotes, we used 1,831 protein sequences of *Encephalitozoon romaleae* SJ-2008 (TAXID: 1178016), and for prokaryotes, we used 253 protein sequences from *Candidatus Portiera aleyrodidarum* BT-B-HRs (TAXID: 1206109).

All taxonomic names and their taxon identifiers (TaxIDs) of the organisms in this study are based on NCBI taxonomy database (39). They are listed in Dataset S1 of *SI Appendix*, where "N/A" indicates an unassigned taxonomic order.

FFP Method. The method (16) and two examples of the application of the method (20, 21) have been published. A brief summary of the two steps taken specifically for this study is described below.

In the first step, we describe the proteome sequence of each organism by the collection of all unique *n*-grams (features), which are short peptide fragments, generated by a sliding "window" of 13 amino acids wide, along the whole-proteome sequence of the organism (see *Choice of the "Best" Descriptor and "Optimal" Feature Length for a Whole-Proteome ToL* below). Some features may be present more than once, so we log the counts. The collected *n*-grams contain the complete information to reconstruct the whole-proteome sequence of the organism. Then, since each organism's proteome has a different size, we convert all the counts of features to frequencies by dividing by the total number of counts for each proteome. Thus, now, each organism is represented by the FFP of its proteome sequence.

The second step is comparing the two FFPs to measure the degree of difference ("divergence") between the two FFPs by JSD (see next section), which measures the degree of difference between two proteome sequences, that is, two FFPs in this study. All pairwise JSDs, then, form a divergence distance matrix, from which we construct the whole-proteome ToL and calculate all of the branch lengths (see below).

JSD and Cumulative Branch Length. JSD (19) values are bound between 0 and 1, corresponding to the JSD value between two FFPs of identical proteome sequences and two completely different proteome sequences, respectively. Any amino acid differences caused by genomic point substitutions, indels, inversion, recombination, loss/gain of genes, etc., as well as other unknown mechanisms, will bring JSD somewhere between 0.0 and 1.0, depending on the degree of information divergence. In this study, the collection of the JSDs for all pairs of extant organisms plus four outgroup members constitute the divergence distance matrix. BIONJ (40, 41) is used to construct the whole-proteome ToL. For convenience of comparison and visibility, all branch lengths are multiplied by 200. This scaling brings CGD from the root to the leaf node of an extant organism to 100 on average, corresponding to the fully evolved genomic divergence of the organism.

A CGD of an internal node is defined as the cumulative sum of all of the scaled branch lengths from root to the node along the presumed evolutionary lineage of the node. (Unscaled JSD values may differ among the JSDs of whole genome, whole transcriptome, and whole proteome. However, since the latter two are derived from the whole-genome sequence, the scaled CGD values of all three are expected to be the same or very similar.)

Choice of the "Best" Descriptor and "Optimal" Feature Length for a Whole-Proteome ToL. For the purpose of constructing the most stable ToL, two key decisions to be made are the choice of the descriptor for the whole-genome information system (DNA sequence of genome, RNA sequence of transcriptome, or amino acid sequence of proteome) and the choice of the optimal feature length of FFP to calculate the "divergence distance" between a pair of FFPs. Since there is no a priori criteria to guide the making of the choices (for other choices, see ref. 42), we took an empirical approach, learned from our earlier studies of building whole-genome trees for the kingdoms of prokaryotes and fungi (20, 21), where we took an operational criterion that the best choice should produce the most topologically stable ToL, as measured by Robinson–Foulds (R-F) metric (43) in PHYLIP package (44), which estimates the topological difference between two ToLs, one with optimal feature length of l and the other with l + 1. The results of the search showed (*SI Appendix*, Fig. S1) that, among the three types of the ToLs, the whole-proteome sequence-based ToL is most topologically stable because it converges to the ToL with lowest R-F metric (near zero) and remains so for largest range of feature length starting from feature length of about 12. In this study, we use l = 13 for the optimal feature length. As for the physical meaning of the optimal feature length, we can infer it from the experiment with books without spaces and delimiters (16), where it approximately corresponds to the feature length at which the number of "vocabulary," the features with unique sequences, is the maximum among all books compared (16). For the optimization criteria different from "the most stable ToL," the best descriptor and optimal feature length may vary depending on the information type and size as well as genomic features important for meeting the criteria, such as noncoding regions, organelle information, and others.

"Outgroup" Members. For the outgroup of our study, we used the shuffled proteome sequences (45, 46) of two eukaryotic and two prokaryotic organisms as in our earlier study of fungi (21): For prokaryotes, we chose *Candidatus Portiera aleyrodidarum* BT-B-Hrs (Gram-negative proteobacteria) with the smallest proteome size of 253 proteins and *Ktedonobacter racemifer* DSM 44,963 (green nonsulfur bacteria) with the largest proteome size of 11,288 proteins; for eukaryotes, we chose two fungi: a Microsporidia, *E. romaleae* SJ-2008, with the smallest proteome size of 1,831 proteins, and a Basidiomycota, *Sphaerobolus stellatus*, with the largest proteome size of 35,274 proteins.

Computer Code Availability. The FFP programs for this study (2v.2.1) written in GCC(g++) are available in GitHub: https://github.com/jaejinchoi/FFP.

Web Address Links. TransDecoder for translating the transcriptome sequence to amino acid sequence is available at https://github.com/TransDecoder/TransDecoder/wiki; treegenesdb at https://treegenesdb.org/; and FTP at https://treegenesdb.org/FTP/.

Hemimastigotes transcriptome data are from https://datadryad.org/stash/dataset/doi:10.5061/dryad.n5g39d7.

ACKNOWLEDGMENTS. We thank Dr. Byung-Ju Kim (formerly at University California, Berkeley, CA, and Yonsei University, Seoul, Korea, and currently at Incheon National University, Incheon, Korea) for numerous discussions and advice on computer programming and algorithm development relevant to this study. We also gratefully acknowledge the expert comments and advice from our colleagues at University of California, Berkeley: Prof. John Taylor on fungi; Profs. James Patton and David Wake on mammals and reptiles; Profs. Alex Glaser and Hiroshi Nikaido on prokarya; Profs. Kip Will and Peter Oboyski on insects; Prof. Rauri Bowie on birds; Prof. Bruce Baldwin on plants; and Prof. Nicole King on protists. Special thanks to Prof. Norman Pace of University of Colorado (Boulder, CO) and Dr. Eugene Koonin of NCBI, National Institutes of Health, for their constructive comments and advice on the various versions of the paper. Rigel Sisson made the silhouettes in Figs. 1 and 2. This research was partly supported by a grant (to S.-H.K.) from World Class University Project, Ministry of Education, Science and Technology, Republic of Korea, and a gift grant to University of California, Berkeley (to support J.C.). S.-H.K. acknowledges having an appointment as Visiting Professorships at Yonsei University, Korea Advanced Institute of Science and Technology, and Incheon National University in South Korea during the manuscript preparation.

1. N. R. Pace, Mapping the tree of life: Progress and prospects. *Microbiol. Mol. Biol. Rev.* **73**, 565–576 (2009).
2. C. R. Woese, G. E. Fox, Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5088–5090 (1977).
3. C.R. Woese, O. Kandler, M.L. Wheelis. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576–4579 (1990)
4. F. D. Ciccarelli *et al*., Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
5. D. Hillis, M. V. Price, R. W. Hill, D. W. Hall, M. J. Laskowski, *Principles of Life*, D. M. Hillis, M. V. Price, R. W. Hill, M. J. Laskowski, D. W. Hall, Eds. (Sinauer Associates and Macmillan Publishers, Sunderland, MA, and New York, ed. 3, 2018).
6. E. Pennisi, Modernizing the tree of life. *Science* **300**, 1692–1697 (2003).
7. C. E. Hinchliff *et al*., Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 12764–12769 (2015).
8. L. A. Hug *et al*., A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
9. C. J. Castelle, J. F. Banfield, Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).

10. S. B. Hedges, S. Kumar, "Discovering the timetree of life" in *Timetree of Life*, S. B. Hedges, S. Kumar, Eds. (Oxford University Press, 2009), pp. 3–18.
11. P. Puigbò, Y. I. Wolf, E. V. Koonin, Search for a "Tree of Life" in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).
12. W. F. Doolittle, Uprooting the tree of life. *Sci. Am.* **282**, 90–95 (2000).
13. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: The phylogenetic jury is still out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).
14. S. Claude, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
15. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990).
16. G. E. Sims, S. R. Jun, G. A. Wu, S. H. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 2677–2682 (2009).
17. W. B. Cavnar, J. M. Trenkle, A. A. Mi, "N-gram-based text categorization," in *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. https://www.semanticscholar.org/paper/N-gram-based-text-categorization-Cavnar-Trenkle/1c610a7e67b578de78436e8959b3ea462ca3e56d. Accessed 30 January 2020.
18. K. Hornik, J. Rauch, C. Buchta, I. Feinerer, textcat: n-Gram based text categorization. R package version 1.0-0 (2013). http://CRAN.R-project.org/package=textcat. Accessed 26 January 2020.
19. J. Lin, Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
20. S.-R. Jun, G. E. Sims, G. A. Wu, S.-H. Kim, Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 133–138 (2010).
21. J. Choi, S.-H. Kim, A genome Tree of Life for the Fungi kingdom. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9391–9396 (2017).
22. A. Auton *et al.*; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. S. Mallick *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
24. O. Lukjancenko, T. M. Wassenaar, D. W. Ussery, Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708–720 (2010).
25. A. Harish, A. Tunlid, C. G. Kurland, Rooted phylogeny of the three superkingdoms. *Biochimie* **95**, 1593–1604 (2013).
26. A. Harish, C.G. Kurland, Empirical genome evolution models root the tree of life. *Biochimie* **138**, 137–155 (2017)
27. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).
28. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).
29. A. Spang *et al.*, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
30. A. G. B. Simpson, C. Slamovits, J. M. Archibald, "Protist diversity and eukaryote phylogeny" in *Handbook of the Protists*, J. M. Archibald, A. G. B. Simpson, C. Slamovits, Eds. (Springer, ed. 2, 2017).
31. F. U. Battistuzzi, S. B. Hedges, A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).
32. G. Lax *et al.*, Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
33. P. Simion *et al.*, A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**, 958–967 (2017).
34. E. V Koonin, The biological Big Bang model for the major transitions in evolution. *Biol. Direct* **2**, 21 (2007).
35. A. Rokas, D. Krüger, S. B. Carroll, Animal evolution and the molecular signature of radiations compressed in time. *Science* **310**, 1933–1938 (2005).
36. N. Eldredge, S. J. Gould, *Models in Paleobiology*, T. F. Schopf, Ed. (Cooper and Co., San Francisco, 1972), pp. 82–115.
37. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
38. J. L. Wegrzyn, J. M. Lee, B. R. Tearse, D. B. Neale, TreeGenes: A forest tree genome database. *Int. J. Plant Genomics* **2008**, 412875 (2008).
39. E. W. Sayers *et al.*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **40**, D13–D25 (2012).
40. N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
41. O. Gascuel, BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
42. A. Zielezinski *et al.*, Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, 144 (2019)
43. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
44. J. Felenstein, PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 163–166 (1989).
45. D. E. Knuth, "Seminumerical algorithms" in *The Art of Computer Programming* (Addison-Wesley, Boston, ed. 3, 1973).
46. R. A. Fisher, F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (Oliver and Boyd, London, 1948).
47. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).